Peer review and citation data in predicting university rankings, a large-scale analysis

Peer review and citation data in predicting university rankings, a large-scale analysis

David Pride and Petr Knoth

The Knowledge Media Institute, The Open University, Milton Keynes, UK.
{david.pride, petr.knoth}@open.ac.uk

Abstract. Most Performance-based Research Funding Systems (PRFS) draw on peer review and bibliometric indicators, two different methodologies which are sometimes combined. A common argument against the use of indicators in such research evaluation exercises is their low correlation at the article level with peer review judgments. In this study, we analyse 191,000 papers from 154 higher education institutes which were peer reviewed in a national research evaluation exercise. We combine these data with 6.95 million citations to the original papers. We show that when citation-based indicators are applied at the institutional or departmental level, rather than at the level of individual papers, surprisingly large correlations with peer review judgments can be observed, up to $r <= 0.802$, $n = 37$, $p < 0.001$ for some disciplines. In our evaluation of ranking prediction performance based on citation data, we show we can reduce the mean rank prediction error by 25% compared to previous work. This suggests that citation-based indicators are sufficiently aligned with peer review results at the institutional level to be used to lessen the overall burden of peer review on national evaluation exercises leading to considerable cost savings.

1 Introduction

Since the late 20th century there has been a seismic shift in many countries in how research is funded. In addition to traditional grant or patronage funding, there is growing use of Performance-based Research Funding Systems (PRFS) in many countries. These systems fall largely into two categories; those that focus on peer review judgments for evaluation and those that use a bibliometric approach. The UK and New Zealand both have systems heavily weighted towards peer review. Northern European countries other than the UK tend to favour bibliometric methodologies whereas Italy and Spain consider both peer review judgments and bibliometrics. Research Evaluation Systems overall have dual and potentially dichotomous ends, firstly identifying the best quality research but also, in many cases, the distribution of research funds. There is, however, a large variance in the level of institutional funding granted based on the results of these exercises. The UK's Research Councils distribute £1.6 billion annually entirely on the basis of the results of the Research Excellence Framework (REF) which is the largest single component of university funding. At the other end

of the scale, the distribution of funds based on the results of the Finnish PRFS is just 3% of the total research budget. Furthermore, the PRFS in Norway and Australia are both used for research evaluation but are not used for funding distribution [1]. Peer-review based PRFS are hugely time-consuming and costly to conduct. In this investigation we ask how well do the results of peer-review based PRFS correlate with bibliometric indicators at the institutional or disciplinary level. A strong correlation would indicate that metrics, where available, can lessen the burden of peer review on national PRFS leading to considerable cost savings, while a weak correlation would suggest each methodology provides different insights.

To our knowledge, this is the first large-scale study exploring the relationship between peer-review judgments and citation data at the institutional level. Our study is based on a new dataset compiled from 190,628 academic papers in 36 disciplines submitted to UK REF 2014, article level bibliometric indicators (6.95m citations) and institutional / discipline level peer-review judgments. This study demonstrates that there is a surprisingly strong correlation between an institutions' *Grade Point Average* (GPA) ranking for its outputs submitted to the UK Research Excellence Framework for many Units of Assessment (UoAs) and citation data. We also shows that this makes it possible to predict institutional rankings with a degree of accuracy in highly cited disciplines.

2 Related work

There has long been wide ranging and often contentious discussion regarding the efficacy of both peer review and bibliometrics and whether one or other, or both should be used for Research Evaluation. Several other studies have specifically investigated the correlation between the results of different nations' peer review focused Performance-based Research Funding Systems and bibliometric indicators. Anderson [2] finds only weak to moderate correlation with results from the New Zealand PRFS and a range of traditional journal rankings. The highest correlation is $r = 0.48$ with the Thomson Reuters Journal Citation Report. However Anderson states that this may be due to the much broader scope of research considered by PRFS processes and the additional quality-related information available to panels. Contrary to Anderson, Smith [3] used citations from Google Scholar (GS) and correlated these against the results from the New Zealand PRFS in 2008. He found strong correlation, $r = 0.85$ for overall PRFS results against Google Scholar citation count.

A comprehensive global PRFS analysis was conducted by Hicks in 2012. Hicks states there is convincing evidence that when PRFS are used to define league tables this creates powerful incentives for institutions to attempt to 'game' the process, whether in regards to submission selection or staff retention and recruitment policies [1]. A UK government funded report, The Metric Tide, was published in 2015 and gave a range of recommendations for the use of metrics in research evaluation exercises. The Metric Tide study had access to the anonymised scores for the individual submissions to the REF and was therefore

directly able to compare on a paper by paper basis the accuracy of a range of bibliometric indicators. This study tested correlations with a range of different bibliometric measures and found correlation with rankings for REF 4* and 3* outputs for some UoAs. Metrics found to have moderately strong correlations with REF scores for a wide range of UoAs included: number of tweets; number of Google Scholar citations; source normalised impact per paper; SCImago journal rank and citation count [4].

However, The Metric Tide study used different citation metrics and citation data sources from our approach. It is at the institutional UoA level that our study reveals some of the strongest correlations, higher than previously shown. In a related study, Mryglod et al. [5] used departmental h-index aggregation to predict REF rankings. Their work was completed before December 2014 when the REF results were published and contained ranking predictions based on their model with some degree of success. They also experimented by normalising the h-index for each year between 2008 and 2014 but surprisingly found little evidence that timescale played a part in the strength of the correlations they found. An ad hoc study by Bishop [6] also found a moderate to strong correlation between departmental research funding based on the results of the UK's Research Assessment and Evaluation (RAE) exercise conducted in 2008, and departmental h-index. Mingers [7] recently completed an investigation that collected total citation counts from Google Scholar (GS) for the top 50 academics[1] from each UK institute and he found strong correlations with overall REF rankings. To our knowledge, ours is the first large-scale in-depth study that investigates the correlation between citation data and peer review rankings by discipline at the institutional level, taking into account all papers submitted to REF.

3 Results

For this study we used data from the UK's Research Excellence Framework (REF). The last REF exercise undertaken in the UK in 2014 was the largest overall assessment of universities' research output ever undertaken globally. These experiments focus on the academic outputs (research papers) component of the REF, for which the metadata are available for download from the REF website. The REF 2014 exercise peer reviewed and graded approximately 191,000 outputs from 154 institutions and in 36 Units of Assessment (UoAs) from zero to four stars. The grading for each submission was determined according to *originality, significance and rigour*. The peer review grades for the individual submissions were aggregated for each UoA to produce a *Grade Point Average* for each institute. The rankings are of critical importance to the institutes as approximately £1.6 billion in QR funding from central government is distributed annually entirely on the basis of the REF results [8].

Each of the REF peer review panels individually chose whether or not to use citation data to inform their decisions. Eleven out of 36 selected to do so and were

[1] If there were not 50 academics then the total number of academics on GS for that institute was used.

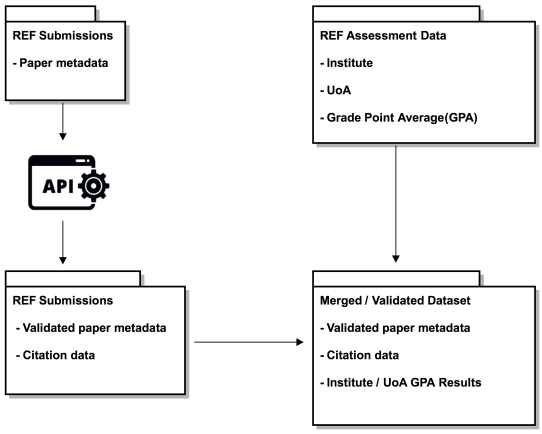

Fig. 1. Citation enrichment workflow used in dataset creation.

provided with citation data from Elsevier Scopus to assist their decision making. For each area and age of publication they were given the number of citations required to put the paper in the top 1%, 5%, 10% or 25% of papers within its area. This gave REF reviewers a subject-level benchmark against which to consider the citation data.[9]

Whereas the aggregate GPA ranking for all UoAs and all institutes is publicly available, it is now not possible to obtain a direct comparison between citation data and the individual rankings for each submission as HEFCE state that these data were destroyed. The rationale behind this was to preempt any requests for this data under the Freedom of Information Act. [10].

3.1 Dataset

The dataset creation procedure is depicted in Figure 1. We first downloaded the REF 2014 submission list [9]. For each output, the list contains; publication title, publication year, publication venue, name of institute and UoA. These fields were fully populated for 190,628 out of 190,963 submissions to the 'outputs' category of the REF process.

We decided to utilise the Microsoft Academic Graph (MAG) to enrich the REF submission list with citation information. At the time of the experiment MAG contained approximately 168m individual papers and 1.15 billion citation pairs. This decision was motivated by the fact that while Scopus, operated by Elsevier, was used to provide citation data to the REF process, the free version of the Scopus API service is limited to 20,000 requests per week. It would have therefore taken almost two months to gather the required data which was not practical as this was more than 10 times slower than using MAG. Additionally, studies by [11] and [12] have recently confirmed how comprehensive the MAG

UoA / Subject	Outputs	% in MAG	Citations	MCPP
Public Health	4,881	94.61%	505,950	109.56
Clinical Medicine	13,394	90.78%	1,278,810	105.17
Physics	6,446	84.51%	491,151	90.15
Biological Sciences	8,608	92.20%	620,009	78.12
Earth Systems / Environment	5,249	91.64%	315,429	65.58
Chemistry	4,698	87.71%	246,361	59.78
Allied Health Professions	10,358	89.35%	402,033	43.43
Ag. Vet. and Food Science	3,919	90.76%	150,959	42.44
Comp. Science and Informatics	7,645	89.22%	284,815	41.76
Economics and Econometrics	2,600	88.81%	95,591	41.4

Table 1. UoAs with the highest mean citations per paper (MCPP).

Number of Units of Assessment (UoAs)	36
Number of institutes	154
Number of UoAs/institution pairs	1,911
Number of submissions (papers)	190,628
Number of submissions (papers) in MAG	145,415
Number of citations	6,959,629

Table 2. Dataset statistics

citation data are. We could not utilise Google Scholar as it does not offer an API and prohibits 'scraping' of data.

We systematically queried the MAG Evaluate API for each submission using a normalised version of the publication's title (lower case, diacritics removed). This returned a set of MAG IDs which which were potential matches of the article. We subsequently queried the MAG Graph Search API to validate each of the potential matches. We accepted as a match the most similar publication title that had at least 0.9 cosine similarity. This threshold was set by manually observing about one hundred matches. Using this process we successfully matched 145,415 REF submissions with 6.95 million citations, corresponding to a recall of 76% of the total initial REF submission list.

Table 1 is ordered by the mean citations per paper (MCPP) and shows total number of submissions, percentage of these submissions available in MAG and the total citations of these submissions.

Additionally, as described in Figure 1, we downloaded the Assessment Data from the REF 2014 website. These data contain the GPA, calculated by aggregating the peer review assessment results of individual papers for each given institution per UoA. We then joined these data with the enriched REF submission list by institution name and UoA. By doing so, we obtained 1,911 UoA/institution pairs together with their peer assessment information (GPA) and corresponding lists of submissions and their citation data (Table 2).

The full dataset used in our experiments and all results can be downloaded from Figshare. [2]

[2] https://figshare.com/s/69199811238dcb4ca987

	UoA	mn_{2017}	med_{2017}	mn_{2014}	med_{2014}	cd
1	Chemistry	0.663	**0.802**	0.637	0.738	Y
2	Biological Sciences	0.188	**0.797**	0.288	0.785	Y
3	Aero. Mech. Chem. Engineering	**0.771**	0.758	0.745	0.760	N
4	Social Work and Policy	0.697	**0.752**	0.629	0.635	N
5	Comp. Sci. and Informatics	0.715	**0.743**	0.720	0.678	Y
6	Economics	0.750	0.737	0.760	**0.770**	Y
7	Earth Systems and Enviro. Sciences	0.472	**0.707**	0.512	0.686	Y
8	Clinical Medicine	0.654	**0.677**	0.666	0.662	Y
9	Public Health and Primary Care	0.535	**0.674**	0.607	0.653	Y
10	Physics	0.600	**0.666**	0.627	0.605	Y
...						
27	Comm. Cultural and Media Studies	**0.369**	0.355	0.334	0.267	N
28	Philosophy	0.352	**0.353**	0.268	0.270	N
29	Law	0.318	0.159	**0.365**	0.136	N
30	Theology and Religious Studies	0.404	0.154	**0.439**	0.153	N
31	English Language and Literature	-0.168	**0.102**	-0.192	0.094	N
32	Art and Design	0.157	0.075	**0.187**	0.118	N
33	Anthropology and Dev. Studies	0.062	-0.009	**0.222**	0.145	N
34	Modern Languages and Linguistics	0.141	-0.069	**0.182**	0.188	N
35	Classics	0.155	-0.07	0.079	**0.285**	N
36	Music Drama Dance & Perf. Arts	0.046	-0.094	0.051	**0.039**	N

Table 3. Correlation between REF GPA output rankings and citation data

3.2 How well do peer review judgments correlate with citation data at the institutional level?

Once we assembled the full dataset, we extracted the following overall citation statistics: mean citations in December 2017 (mn_{2017}), median citations in December 2017 (med_{2017}), mean citations at the time of the REF exercise (mn_{2014}), and median citations at the same point (med_{2014}). These data were then used to test the correlation between citation data and REF GPA rankings for outputs for every institute in every UOA. The ten highest and ten lowest measured correlations by UoA are shown in Table 3. The citation data (cd) column denotes whether the REF judging panels considered citation data in their deliberations. While we attempted to run correlations with other similar aggregate functions, these are not shown in this table as they have far lower correlations with GPA.

Strong positive correlations can be observed at the discipline level for a large proportion of the UoAs, particularly for median citation count in 2017. Whilst the correlation was most often stronger for those UoAs that had used citation data in the REF peer review process, this was not always the case. Aeronautical and Mechanical engineering and Social work & Policy are two disciplines, which did not use citation data yet, show very strong correlations with GPA results. At the lower end of the scale, there was little correlation between GPA ranking and citation data, notably for those subjects covered by REF panels C and D.[4]. Lack of coverage in many of these areas is, however, understandable as these are disciplines which do not always produce journal articles, conference

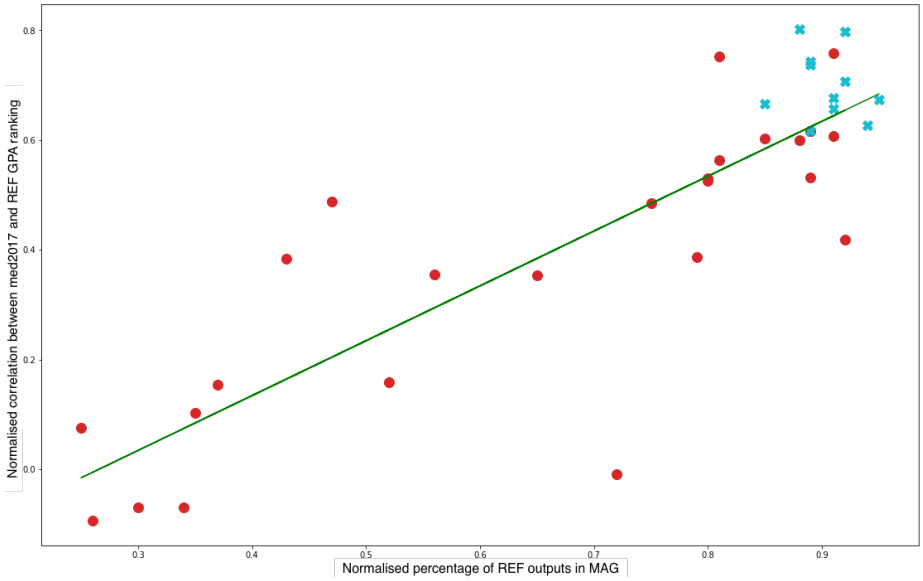

Fig. 2. Correlation between med_{2017} citations per UoA and GPA against the coverage of REF submissions in MAG for all UoAs. An 'o' represents a non-citation based UoA whilst and 'x' denotes a UoA that used citations.

proceedings and other digitally published and highly citable artifacts as their main type of output. There is, however, clear delineation between the more highly correlated UoAs and those less correlated. The UoAs with the lowest are distinct from the rest, they are having a very weak or no correlation ($r <= 0.159, n = 37, p < 0.001$). Those above this level have a medium to strong correlation ($r > 0.353, n = 37, p < 0.001$). The low correlation for mean citations for Biological Sciences is explained by a single paper which was the most highly cited paper in the UoA. This paper received 4,626 citations, 58% more than next cited paper and nine times as many citations as all other submissions for that institute combined. Furthermore, this paper came from second lowest ranked (by GPA) of 44 institutes. Had this paper been discounted from the correlations, the prediction results would have been far more clearly aligned with the other UoAs (mn2017=0.782, mn2014=0.766).

The variance of citation data coverage across UoAs led us to explore whether there could be a relationship between the strength of the correlations GPA and citation data correlation with the coverage of citation in a given UoA. Figure 2 plots this for both the UoAs that used citation data and those that did not. While the graph confirms that the highly cited UoAs in MAG are those UoAs that used citation data, it indicates that a few UoAs that did not also exhibit strong correlations. Unsurprisingly, the plot suggests that there might be a small bias exhibited by extra correlation strength in UoAs that utilised citation data. However, given the small number of UoAs, this is not statistically significant.

REF UoA / Rank	**GPA**	med_{2017}	**mc2017**	**rdiff**	med_{2014}	**mc2014**	**rdiff**
Chemistry							
Liverpool	3.44	Liverpool	64	**0**	Liverpool	26	**0**
Cambridge	3.42	Cambridge	54	**0**	Lancaster	25	+8
Oxford	3.32	Warwick	53	+3	Oxford	22	**0**
UEA	3.29	Bath	51	+12	Cambridge	22	-2
Bristol	3.26	Oxford	50	-2	Queen Mary	20	+2
Bio Sciences							
ICR	3.44	ICR	77	**0**	ICR	31	**0**
Newcastle	3.33	Queen Mary	66	+15	Sheffield	26	+5
Dundee	3.3	Imperial	59	+1	Imperial	25	+1
Imperial	3.26	Sheffield	56	+3	Leeds	24	+27
Oxford	3.26	Edinburgh	55	+4	Edinburgh	23	+4
Aero. Mech.							
Cambridge	3.34	Cambridge	25	**0**	Cambridge	9	**0**
Imperial	3.12	Imperial	23	**0**	Imperial	8	**0**
UCL	3.06	Sheffield	19	+2	Brighton	7	+13
Cranfield	3.01	Brighton	18	+12	Manchester	6	+4
Sheffield	3.01	Manchester	17	+3	Sheffield	6	0

Table 4. Rankings by GPA and predictions produced using med_{2017} and med_{2014} respectively for the three most highly correlated UoAs.

3.3 How well can citation data predict peer review based institutional rankings?

Tables 4 shows the top five institutions for Chemistry, Biological Sciences and Aeronautical and Mechanical Engineering as ranked in the REF by GPA and predictions of ranking using med_{2017} and med_{2014} respectively. *mc2017* and *mc2014* show the median citation count for that institute. *Rdiff* shows the rank difference when ranked by a particular citation metric. The prediction performance indicated in these tables is not unique, in four of the five top UoAs by correlation strength the highest ranked institute is predicted correctly by both med_{2014} and med_{2017}.

Table 5 demonstrates the effectiveness of predicting based on med_{2014} for the 10 most highly cited UoAs. To compare the prediction error, expressed by *rdiff*, across UoAs, we calculated the mean rank difference normalised by number of institutions (*nrdiff*). To express overall prediction accuracy, we used Mean Average Precision (MAP). The HEI column denotes the number of institutes submitting to that UoA. The parameter rt denotes the prediction rank tolerance. For example, $rt = 3$ indicates that a prediction within 3 positions of the original assessment result will be considered as correct. Given the simplicity of the prediction method, this is a strong indication of the power of citation data in this task. One could reasonably expect that further improvements can be made by employing more sophisticated indicators. However, as the predictions are not as good for UoAs that have lower than average mean citations per paper , we would restrain from recommending the use of citation data unaccompanied by peer review assessments in those UoAs.

UoA	HEIs	*rdiff*	*nrdiff*	*MAP rt=3*	*MAP rt=5*	*MAP rt=10*	*MAP rt=10%*	*MAP rt=20%*	*MAP rt=30%*
Comp Sci.	89	12.39	0.139	0.19	0.32	0.50	0.46	0.75	0.87
Ag. Vet.	29	4.02	0.139	0.45	0.65	0.86	0.45	0.68	0.86
Clinical Med.	31	4.38	0.141	0.51	0.70	0.93	0.51	0.77	0.93
Allied H.	83	12.03	0.145	0.20	0.30	0.55	0.43	0.72	0.86
Economics	28	4.07	0.145	0.57	0.71	0.92	0.57	0.78	0.92
Chemistry	37	5.51	0.149	0.54	0.56	0.83	0.54	0.78	0.86
Earth Systems	45	7.24	0.161	0.40	0.51	0.77	0.51	0.68	0.84
Public Health	32	5.18	0.162	0.50	0.62	0.84	0.50	0.68	0.84
Bio. Science	44	7.59	0.173	0.34	0.52	0.72	0.52	0.66	0.79
Physics	41	7.36	0.180	0.36	0.53	0.78	0.43	0.73	0.80
All (mean)	**45**	**6.98**	**0.153**	**0.41**	**0.54**	**0.77**	**0.49**	**0.72**	**0.86**

Table 5. Rank prediction quality for top 10 UoAs with the highest mean citations per paper.

UoA	HEIs	*rdiff*	*nrdiff*	*MAP rt=3*	*MAP rt=5*	*MAP rt=10*	*MAP rt=10%*	*MAP rt=20%*	*MAP rt=30%*
Mryglod [5]									
Chemistry	29	4.89	0.169	0.37	0.82	0.82	0.37	0.82	0.82
Physics	32	8.63	0.270	0.28	0.40	0.65	0.28	0.46	0.65
Bio Science	31	8.38	0.270	0.22	0.38	0.70	0.22	0.51	0.64
All (mean)	**31**	**7.30**	**0.24**	**0.29**	**0.53**	**0.72**	**0.29**	**0.60**	**0.70**
Pride & Knoth (this study)									
Chemistry	29	4.00	0.138	0.68	0.72	0.89	0.68	0.72	0.86
Physics	32	5.68	0.178	0.34	0.59	0.90	0.34	0.75	0.90
Bio Science	31	7.16	0.231	0.35	0.45	0.74	0.35	0.51	0.71
All (mean)	**31**	**5.61**	**0.18**	**0.46**	**0.59**	**0.84**	**0.46**	**0.66**	**0.82**
Improvement		**23%**	**25%**	**59%**	**11%**	**17%**	**59%**	**10%**	**17%**

Table 6. Comparison of the prediction performance of our study with Mryglod et al.[5]

We wanted to compare our prediction performance to the study of Mryglod et al. [5]. In order to conduct a fair and exact comparison, it was necessary to parse a number of institutions from our input data. Mryglod et al. reported they were unable to obtain citation indicators for all institutions in a given UoA. Their study covered three of the top ten highly cited UoAs, we show in Table 6 that our predictions are significantly better across all categories.

4 Discussion

It has been shown in [4], [13] and that many bibliometric indicators show little correlation with peer review judgments at the article level. This study, and those by [7], [2] and [3], demonstrate that some bibliometric measures can offer a surprisingly high degree of accuracy when used at the institutional or departmental level. Our work has been conducted on a significantly larger dataset and our prediction accuracy is higher than shown in previous studies, despite deliberately using fairly simplistic indicators. Several studies including The Metric

Tide [4], The Stern Report [14] and the HEFCE pilot study [15] all state that metrics should be used as an additional component in research evaluation, with peer review remaining as the central pillar. Yet, peer review has been shown by [16], [17] and [18] amongst others to exhibit many forms of bias including institutional bias, gender / age related bias and bias against interdisciplinary research. In an examination of one of the most critical forms of bias, that of publication bias, Emerson [19] noted that reviewers were much more likely to recommend papers demonstrating positive results over those that demonstrated null or negative results.

All of the above biases exist even when peer review is carried out to the highest international standards. There were close to 1,000 peer review experts recruited by the REF, however the sheer volume of outputs requiring review calls into question the exactitude of the whole process. As an example the REF panel for UoA 9, Physics, consisted of 20 members. The total number of outputs submitted for this UoA was 6,446. Each paper is required to be read by two referees. This increases the overall total requirement to read 12,892 paper instances. Therefore each panel member was required to review, to international standards, an average of 644 papers in a little over ten months. If every panel member, worked every day for ten months, each member would need to read and review 2.14 papers *per day* to complete the work on time. This is, of course, in addition to the panelist's usual full-time work load. Moreover, Physics is not an unusual example and many other UoAs tell a similar story in terms of the average number of papers each panel member was expected to review; Business and Management Studies (1,017 papers), General Engineering (868 papers), Clinical Medicine (765 papers). The burden placed on the expert reviewers during the REF process was onerous in the extreme. Coles [20] calculated a very similar figure of 2 papers per day, based on an estimate before the data we now have was available. 'It is blindingly obvious,' he concluded, 'that whatever the panels do will not be a thorough peer review of each paper, equivalent to refereeing it for publication in a journal'. Sayer [21] is equally disparaging in regards to the volume of papers each reviewer was required to read and also expresses significant doubts about the level of expertise within the review panels themselves.

In addition to the potential pitfalls in the current methodologies, there is also the enormous cost to be considered. This was estimated to be £66m for the UK's original PRFS, the Research Assessment Exercise (RAE) in 2008. This rose markedly to £246m for the 2014 Research Excellence Framework. This is comprised of £232M in costs to the higher education institutes and around £14M in costs for the four UK higher education funding bodies. The cost to the institutions was approximately £212M for preparing the REF submissions for the three areas; outputs, impact and environment, with the cost for preparing the outputs being the majority share of this amount. Additionally, there were costs of around £19M for panelists' time. [22]. If bibliometric indicators can in any way lessen the financial burden of these exercises on the institutions this is a strong argument in favour of their usage.

5 Conclusion

This work constitutes the largest quantitative analysis of the relationship between peer reviews (190,628 paper submissions) and citation data (6.9m citation pairs) at an institutional level. Firstly, our results show that citation data exhibit strong correlations with peer review judgments when considered at the institutional level and within a given discipline. These correlations tend to be higher in disciplines with high mean citations per paper. Secondly, we demonstrate that we can utilise citation data to predict top ranked institutions with a surprisingly high precision. In the ten UoAs with the highest number of mean citations per paper we achieve 0.77 MAP with prediction rank tolerance 10 with respect to the REF 2014 results. In four out of five top UoAs by correlation strength, the highest ranked institute in the REF results was predicted correctly. It is important to note that these predictions are based on citation data that were available at the time of the REF exercise.

While our analysis does not answer whether using citation-based indicators we can predict institutional rankings better than by relying on a peer review system, our results evidence that the REF peer review process led to highly similar results as those that could have been predicted automatically using citation data. The 11 REF UoAs with the highest mean citations per paper in MAG are the identical UoAs in which the peer review panels used citation data to inform their decisions. We argue that if peer-review is conducted in the way it was conducted in the REF, then it would have been more cost effective to save a significant proportion of the £246m spent on organising the peer review process [22] and carry out the institutional evaluation purely using citation data, particularly in UoAs with high mean citations per paper.

This has wide implication for PRFS globally. The countries whose PRFS still have a peer review component should carefully consider the way in which the peer review process is conducted. Thus ensuring that the peer review results add a new dimension to the information over that which can be obtained by predictions based on citation data alone. However, this advice only applies when the goal of the PRFS is to rank institutions, as it is the case in the UK REF, rather than individual papers or researchers.

6 Acknowledgements

This work has been funded by Jisc and has also received support from the scholarly communications use case of the EU OpenMinTeD project under the H2020-EINFRA-2014-2 call, Project ID: 654021

References

1. Hicks D. Performance-based university research funding systems. Research policy. 2012;41(2):251–261.

2. Anderson DL, Smart W, Tressler J. Evaluating research–peer review team assessment and journal based bibliographic measures: New Zealand PBRF research output scores in 2006. New Zealand Economic Papers. 2013;47(2):140–157.
3. Smith AG. Benchmarking Google Scholar with the New Zealand PBRF research assessment exercise. Scientometrics. 2008;74(2):309–316.
4. HEFCE. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management; 2015. Available from: http://www.hefce.ac.uk/pubs/rereports/year/2015/metrictide/.
5. Mryglod O, Kenna R, Holovatch Y, Berche B. Predicting results of the Research Excellence Framework using departmental h-index. Scientometrics. 2015 Mar;102(3):2165–2180. Available from: https://doi.org/10.1007/s11192-014-1512-3.
6. Bishop D. An alternative to REF2014?; 2013. Available from: http://deevybee.blogspot.co.uk/2013/01/an-alternative-to-ref2014.html.
7. Mingers J, O'Hanley JR, Okunola M. Using Google Scholar institutional level data to evaluate the quality of university research. Scientometrics. 2017;113(3):1627–1643.
8. HEFCE. Research Excellence Framework 2014: Overview report by Main Panel A and Sub-panels 1 to 6; 2015. Available from: http://www.ref.ac.uk/2014/media/ref/content/expanel/member/Main%20Panel%20A%20overview%20report.pdf.
9. HEFCE. Research Excellence Framework - Results and Submissions; 2014. Available from: http://results.ref.ac.uk/Results.
10. HEFCE. Annex A - Summary of additional information about outputs; 2014. Available from: http://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12a.pdf.
11. Herrmannova D, Knoth P. An analysis of the microsoft academic graph. D-Lib Magazine. 2016;22(9/10).
12. Hug SE, Brändle MP. The coverage of Microsoft Academic: Analyzing the publication output of a university. Scientometrics. 2017;113(3):1551–1571.
13. Baccini A, De Nicolao G. Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. Scientometrics. 2016;108(3):1651–1671.
14. Stern N, et al. Building on success and learning from experience: an independent review of the Research Excellence Framework. London: UK Government, Ministry of Universities and Science. 2016;.
15. HEFCE. Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework; 2016.
16. Hojat M, Gonnella JS, Caelleigh AS. Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process. Advances in Health Sciences Education. 2003;8(1):75–96.
17. Lee CJ, Sugimoto CR, Zhang G, Cronin B. Bias in peer review. Journal of the Association for Information Science and Technology. 2013;64(1):2–17.
18. Smith R. Peer review: a flawed process at the heart of science and journals. Journal of the royal society of medicine. 2006;99(4):178–182.
19. Emerson GB, Warme WJ, Wolf FM, Heckman JD, Brand RA, Leopold SS. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. Archives of internal medicine. 2010;170(21):1934–1939.
20. Coles P. The apparatus of research assessment is driven by the academic publishing industry; 2013. Available from: https://bit.ly/2EfNMeV.
21. Sayer D. Rank hypocrisies: The insult of the REF. Sage; 2014.
22. Technopolis. REF Accountability Review: Costs, benefits and burden; 2015.